基于递归神经网络的视频多目标检测技术 *

华　夏 [1]，王新晴 [1]，马昭烨 [1]，王　东 [1,2]，邵发明 [1]

(1. 中国人民解放军陆军工程大学，南京 210007; 2. 南部战区陆军第二工程科研设计所，昆明 650222)

摘　要：针对现有基于大数据和深度学习的目标检测框架难以实现在低功耗移动和嵌入式设备上实时进行视频目标检测的问题，改进了基于深度学习的目标检测框架 SSD，提出一种改进的多目标检测框架 LSTM-SSD，将其专用于交通场景视频多目标检测。将单图像检测框架与递归神经网络 LSTM 网络相结合，形成交织循环卷积结构，通过采用一种 Bottleneck-LSTM 层提炼传播帧间的特征映射实现了网络帧级信息的时序关联，极大降低了网络计算成本；将时间感知信息与改进的动态卡尔曼滤波算法结合起来，实现了对视频中受光照变化、大面积遮挡等强干扰影响目标的追踪识别；实验表明，改进后的 LSTM-SSD 在应对多目标、杂乱背景、光照变化、模糊、大面积遮挡等检测难度较大的情况时，均能获得较好的效果，相比于其他基于深度学习的目标检测框架，各类目标识别的平均准确率提高了 5~16%，平均准确率均值提高了约 4~10%，多目标检测率提高 4~19%，检测帧率达到 43 fps，基本满足实时性的要求。实现了算法精度与运行速率的平衡，取得较好的检测识别效果。

关键词：机器视觉；深度学习；递归神经网络；卡尔曼滤波；视频多目标检测；卷积神经网络

中图分类号：TP391.41　　doi: 10.19734/j.issn.1001-3695.2018.05.0567

Video multi-target detection technology based on recursive neural network

Hua Xia[1], Wang Xinqing[1], Ma Zhaoye[1], Wang Dong[1, 2], Shao Faming[1]

(1. *PLA Army Engineering University*, *Nanjing* 210007, *China*; 2.the *2nd Institute of Engineering Research & Design*, *Southern Theatre Command*, *Kunming* 650222, *China*)

Abstract: Aiming at the problem that the existing target detection framework based on big data and deep learning is difficult to realize real-time video target detection on low-power mobile and embedded devices, this paper improves the target detection framework SSD (single shot multi-box detector) based on deep learning, and puts forward an improved multi-target detection framework LSTM-SSD (long short term memory, LSTM) , which is dedicated to multi-target detection of traffic scenes video. Combining single image detection frame with recursive neural network lstm network to form an interleaved circular convolution structure, the temporal association of network frame-level information is realized by extracting the feature map between propagation frames by adopting a little neck - lstm layer, which greatly reduces the network calculation cost. Combining the time-aware information with the improved dynamic Kalman filtering algorithm, the tracking and identification of the targets which are influenced by strong interference such as light change and large-area occlusion in the video can be realized. Experimental results show that the improved lstm - SSD can achieve good results when dealing with the difficult detection situations such as multi - targets, cluttered background, light changes, fuzziness and large-area occlusion. compared with other target detection frameworks based on deep learning, the average accuracy rate of all kinds of target identification is increased by 5~16 %, the average accuracy rate is increased by 4~10 %, the multi-target detection rate is increased by 4~19 %, and the detection frame rate reaches 43 frames / s, basically meeting the requirements of real-time. The balance between the accuracy of the algorithm and the running speed is achieved, and a good detection and identification effect is achieved.

Key words: machine vision; deep learning; recurrent neural network; Kalman filter; video multi-target detection; convolutional neural network

0　引言

交通场景中的行人、车辆目标检测与识别是目标检测技术的重要分支， 是自动驾驶、机器人以及智能视频监控等研究领域的核心技术，有着重要的研究意义[1]。

深度学习为基于深层人工神经网络的学习方法，基于深度学习的目标检测算法可应用于多种检测场景，综合性强，能够同时检测和识别多类目标，主动性好。各种类型的人工神经网络结构中，深度卷积网络具有强大的特征提取能力，越来越多的用于图像分类的网络结构被提出，不断提升了深度卷积网络在特征提取方面的优势，在图像识别、图像分割、目标检测、场景分类等视觉任务中，取得了非常好的效果[2]。

收稿日期：2018-05-23；**修回日期：**2018-07-30　**基金项目：基金项目：**国家重点研发计划资助项目（2016YFC0802904）；国家自然科学基金资助项目（61671470）；江苏省自然科学基金资助项目（BK20161470）；中国博士后科学基金第 62 批面上资助项目（2017M623423）

作者简介：华夏（1995-），男，硕士研究生，主要研究方向为计算机图形学、机器视觉、数字图像处理（1614118084@qq.com）；王新晴（1963-），男，教授，博导，博士，主要研究方向为机电控制、智能信号处理、机器视觉；马昭烨，男，讲师，主要研究方向为机电控制、智能信号处理、机器视觉；王东，男，讲师，博士，主要研究方向为机电控制、智能信号处理、机器视觉；邵发明，男，讲师，博士研究生，主要研究方向为机电控制、智能信号处理、机器视觉.

SSD，全称 single shot multibox detector[3]，是 Liu Wei 在 ECCV 2016 上提出的一种目标检测算法，截至目前是主要的检测框架之一，相比 Faster RCNN[4]有明显的速度优势，相比 YOLO[5]又有明显的平均准确率均值优势。SSD 具有如下主要特点：从 YOLO 中继承了将检测问题转换为回归的思路，同时一次即可完成网络训练；基于 Faster RCNN 中的 anchor，提出了相似的 prior box；加入基于特征金字塔[6]的检测方式，相当于半个 FPN[6]思路。尽管 SSD 在特定数据集上已经取得了较高的准确率，具有较好的实时性，但是模型的训练过程非常耗时，对训练样本的质和量依赖严重；通过图像的颜色、边缘等信息来检测目标，其对于弱小目标和大面积遮挡目标等缺乏图像信息的目标检测效果不佳；算法检测效率仍然有待提高，以满足装备运行实时性的要求。

本文针对复杂大交通场景下行人、车辆目标检测任务的特点和需求，对传统 SSD 算法进行了以下两点改进：a）将单图像检测框架与递归神经网络-LSTM 网络相结合，形成交织循环卷积结构，通过采用一种 Bottleneck-LSTM 层提炼传播帧间的特征映射实现了网络帧级信息的时序关联，极大降低了网络计算成本; b）将时间感知信息与改进的动态卡尔曼滤波算法结合起来，实现了对视频中受光照变化、大面积遮挡等强干扰影响目标的追踪识别。实验表明，改进后的 LSTM-SSD 在应对多目标、杂乱背景、光照变化、模糊、大面积遮挡等检测难度较大的情况时，均能获得较好的效果

1　基于时间感知特征映射的视频目标检测框架

SSD 采用了特征金字塔结构进行检测，即检测时利用了 conv4-3，conv-7（FC7），conv6-2，conv7-2，conv8_2，conv9_2 这些大小不同的 feature maps，在多个 feature maps 上同时进行 softmax 分类和位置回归，对弱小目标有较好的检测精度[3]，其网络结构如图 1 所示。

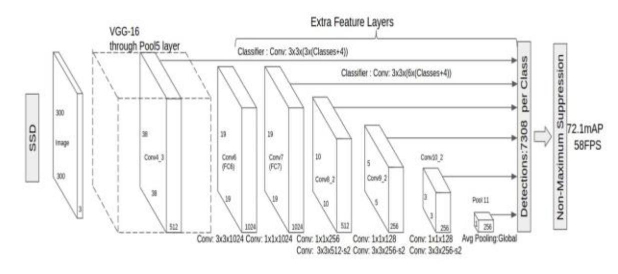


图 1　SSD 网络架构

Fig. 1　SSD network architecture

视频图像数据包含多种时间线索，与单个图像相比，它们可以被展开以获得更准确和稳定的目标检测。由于视频表现出时间连续性，相邻帧中的目标将保持在相似的位置，并且检测结果将基本不变。因此，可以使用来自较早帧的检测结果信息来细化当前帧处的预测结果。由于网络能够跨帧以不同状态检测目标，随着训练时间的推移，网络的预测结果也将变得有更高的置信度，从而有效减少单图像目标检测中存在的不稳定性问题[7]。

最新研究表明，这种连续性可以延伸到特征空间，并且从视频的相邻帧提取的过渡特征映射也具有高度相关相关性。感兴趣的是在特征空间中也添加时间感知机制，而不仅仅是在最终检测结果上添加时间感知机制，这是因为中间层中可用的信息量更大。通过递归网络体系结构将每个帧的特征映射调整到先前帧的相应特征映射上来利用特征级的连续性。

该方法将标准卷积层与卷积 LSTM 网络相结合，生成具有联合卷积递归单元的特征映射。目标是卷积层输出特征映射假设，然后将该假设馈送到 LSTM 并与来自先前帧的时间背景信息融合以输出精细的时间感知特征映射。

人类不会每秒钟都从头开始思考，人类的思想具有连贯性。但是传统的神经网络做不到这一点，这是一个主要的缺点。递归神经网络（recurrent neural network, RNN）较好地解决了这个问题。它们是具有环路的网络，允许信息持续存在。RNN 优势之一在于它们能够将先前的信息关联到当前的任务中，比如用之前的视频帧可以辅助理解当前的视频帧[8]。长短期记忆（long-short term memory, LSTM）是一种特殊的 RNN，它被设计用来避免长期依赖的问题，由于独特的设计结构，LSTM 适合于处理和预测时间序列中间隔和延迟非常长的重要事件。

提出了一种将卷积 LSTMs 结合到单图像检测框架中的方法，作为跨时间传播帧级信息的手段。然而，LSTMs 的简单集成会导致较大的运算量，妨碍网络实时运行。为了解决这个问题，引入了一个 Bottleneck-LSTM[7]，利用它具有深度可分离卷积和 Bottleneck 设计原则的特性，降低计算成本。图 2 是 LSTM-SSD 的网络结构。网络中插入多个卷积 LSTM 层。 每个都以一定的比例传播和提炼特征映射。

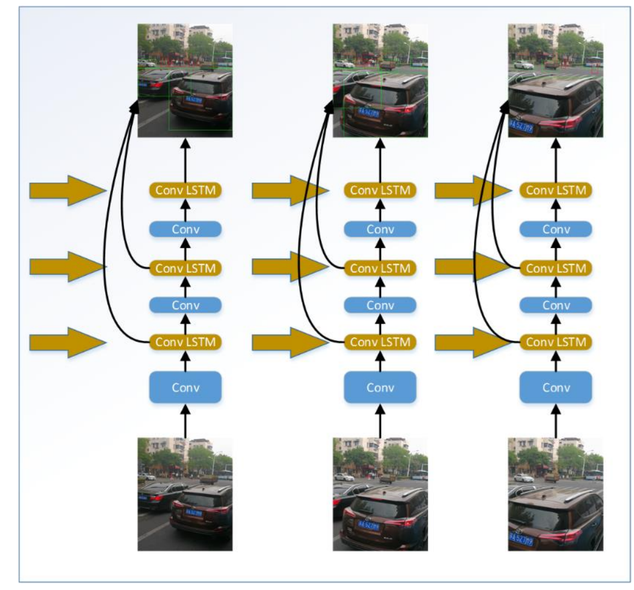


图 2　LSTM 视频目标检测网络框架

Fig. 2　LSTM video target detection network framework

将视频数据视为多帧图像组成的序列，$V=\{I_0,I_1,\cdots,I_n\}$，目标是得到帧级的检测结果$\{D_0,D_1,\cdots,D_n\}$，其中 Dk 表示对图像帧 Ik 的检测结果，包括一系列对各个目标检测的检测框的位置，以及对各个目标的识别置信度。考虑构造一种在线学习机构，使得检测结果 D_k 可以由图像帧 I_{k-1} 进行预测和修正。将预测模型当做函数

$$F(I_t,s_{t-1})=(D_t,s_t) \tag{1}$$

其中：$s_k=\{s_k^0,s_k^1,s_k^2,\cdots,s_k^{m-1}\}$，表示描述视频第 k 帧图像的特征映射向量，构造一个具有 m 层 LSTM 卷积层的神经网络来近似地实现这个函数功能。这个神经网络把特征映射向量 s_{t-1} 中的每个特征映射作为 LSTM 卷积层的输入，可以得到对应的特征映射向量 s_t。要获得整个视频的检测结果，只需通过网络顺序运行每帧图像。

当应用于视频序列时，可以将 LSTM 状态理解为为表示时序的特征。然后，LSTM 可以在每个时序步骤使用时序特征来细化其输入，同时还从输入提取附加的时间信息并更新

其状态。这种精细化模式可以通过在任意中间特征映射上紧接着放置 LSTM 卷积层来应用。特征映射用作 LSTM 的输入，而 LSTM 的输出将在以后的所有计算中替换之前的特征映射。可以将单帧图像目标检测器定义为函数 $G(I_t)=D_t$，该函数将用于构造具有 m 个 LSTM 层的复合网络。可以将这些 LSTM 卷积层看做是将函数 G 的层划分为 m+1 个合适的子网络 $\{g_0,g_1,\cdots,g_m\}$，则

$$G(I_t)=(g_m\circ\cdots g_1\circ g_0)(I_t) \tag{2}$$

○表示哈达玛乘积(hadamard product)。本文同样将任意一层 LSTM 卷积层定义成为函数

$$L_k(M,s_{t-1}^k)=(M_+,s_t^k) \tag{3}$$

其中：M、M_+ 都是同维度的特征映射。则按照时序进行计算，公式如下：

$$\begin{aligned}(M_+^0,s_t^0)&=L_0\left(g_0(I_t),s_{t-1}^0\right)\\(M_+^1,s_t^1)&=L_1\left(g_1(M_+^0),s_{t-1}^1\right)\\&\vdots\\(M_+^{m-1},s_t^{m-1})&=L_{m-1}\left(g_{m-1}(M_+^{m-2}),s_{t-1}^{m-1}\right)\\D_t&=g_m(M_+^{m-1})\end{aligned} \tag{4}$$

图 3 描述了整个模型在处理视频时的输入和输出。实际上， LSTM 层的输入和输出可以具有不同的维度，但是只要每个子网 F 的第一卷积层的输入维度被修改，就可以执行相同的计算。

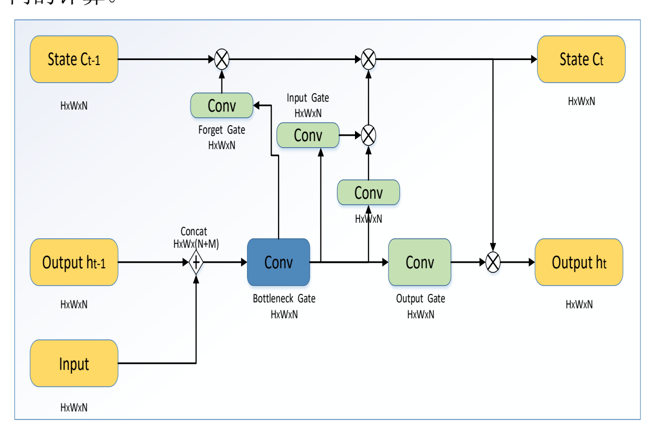


图 3　Bottleneck-LSTM 模型处理视频输入和输出示意图

Fig. 3　Schematic diagram of the Bottleneck-LSTM model processing video input and output

在本文的体系结构中，通过实验选择了 G 的分区。较早地放置 LSTM 会导致较大的数据输入量，并且计算成本爆炸增长导致运算效率低下。为了保证算法的运算效率，仅在具有最低空间维度的特征映射之后考虑 LSTM 放置。

由于需要在单个前向通道中计算多个门，所以 LSTMs 对计算资源有着较高的要求，这极大地影响了网络的整体效率。为了解决这个问题，引入了一系列的更改，使 LSTMs 能够与实时移动目标检测的目的兼容。

首先，考虑调整 LSTM 的维度。通过扩展在文献[7]中定义的通道宽度乘子 α_δ，可以获得对网络结构更好的控制。原始宽度倍增器是用于缩放每个层的通道尺寸的超参数，而不是将这个乘数统一应用于所有层。引入了三个新的参数 α_{base}、α_{ssd}、α_{lstm}，它们控制网络不同部分的信道尺寸。具有 N 个输出通道的基本移动网络中的任何给定层被修改为具有 $N\alpha_{base}$ 个基本输出通道，而 α_{ssd} 应用于所有 SSD 特征映射，α_{lstm} 应用于 LSTM 层。设置 $\alpha_{base}=\alpha$，$\alpha_{ssd}=0.5\alpha$，$\alpha_{lstm}=0.25\alpha$。每个 LSTM 的输出是输入大小的四分之一，这大大减少了所需的计算。

同时通过采用一种新的 Bottleneck-LSTM[7]，极大地提高了传统 LSTM 的运算效率

$$b_t=\phi\left({}^{M+N}W_b^N*[x_t,h_{t-1}]\right) \tag{5}$$

其中：x_t，h_{t-1} 为输入的特征映射，$\phi(x)$=ReLU(x)，ReLU 表示 ReLU 激活。ReLU 表示修正线性单元(Rectified linear unit，ReLU）激活，虽然 ReLU 激活在 LSTMs 中并不常用，但是不改变特征映射的边界很重要，因为 LSTMs 散布在卷积层之间。${}^jW^k*X$ 表示具有权重 W、输入 X、j 输入通道和 k 输出通道的深度可分离卷积。这种修改的好处是双重的。使用瓶颈特征映射减少了门内的计算，在所有实际场景中均优于标准 LSTMs。其次，Bottleneck-LSTM 比标准的 LSTM 更深，而较深的模型优于较宽和较浅的模型[7]。

2　针对受强干扰目标的检测改进策略

复杂交通场景中的遮挡、光照、阴影等强干扰现象会造成目标外观信息损失，致使检测过程中容易出现目标遗漏。训练有素的卷积神经网络可以应对一定程度的干扰，但无法应对大面积遮挡等强干扰造成目标图像信息严重缺失。对此本文提出时空上下文策略，从之前的检测结果中获取有用的先验信息合理预测少量候选区域，增加目标被检测的几率。这一思路借鉴了目标跟踪的方法来优化检测结果[10]。

卡尔曼滤波和粒子滤波常常被用于跟踪算法中。卡尔曼滤波使用有三个前提假设：被建模的系统是线性的；影响测量的噪声属于白噪声；噪声本质上是高斯分布的。很显然，由于摄像机的运动和神经网络本身复杂的非线性映射，目标在视频中的位置和置信度并非线性变化的[11]。但本文只是将滤波作为提高候选区域质量的辅助手段，而且在短时间内目标可以近似看成线性运动。所以本文选择卡尔曼滤波作为在前一帧和当前帧之间传递目标信息的工具，结合目标检测任务设计卡尔曼滤波模型。

$D_k=\{X_k^0,X_k^1,\cdots,X_k^n\}$ 表示使用未加入滤波的检测器对图像帧 I_k 的检测结果， $X_k^t=[x_k^t,y_k^t,a_k^t,b_k^t,c_k^t,d_k^t]$ x, y, a, b ,d 分别为第 k 帧某一目标 t 外接矩形框的左上角坐标和宽高，c 为目标置信度，d 为目标所属类别。通过 LSTM 可以获得视频第 k+1 帧的检测结果 D_{k+1} 的预测值 $\hat{D}'_{k+1}$。但是因为预测过程中存在噪声等因素干扰产生的误差，如果不对预测结果加以修正，那么在视频检测的过程中误差将因为迭代过程而被无限地放大，为了避免出现这种情况，将视频第 k+1 帧的初检测结果 Z_{k+1} 作为测量值对 LSTM 的预测值 $\hat{D}'_{k+1}$ 进行修正，即采用“预测+测量反馈”的方式获得视频第 k+1 帧的检测结果 D_{k+1} 的估计值 $\hat{D}_{k+1}$。则系统的估计值滤波方程为

$$\hat{X}_{k+1}^t=A_k\hat{X}_k^{t\prime}+K_{k+1}\left(Z_{k+1}^t-H_{k+1}A_k\hat{X}_k^{t\prime}\right) \tag{6}$$

系统的测量方程为

$$Z_{k+1}^t=HX_{k+1}^t+v_{k+1} \tag{7}$$

卡尔曼增益方程为

$$K_{k+1}=P_{k+1/k}H^T\left(HP_{k+1/k}H^T+V_{k+1}\right)^{-1} \tag{8}$$

预测误差协方差矩阵方程为

$$P_{k+1/k} = AP_k A^T + W_k \tag{9}$$

修正误差协方差矩阵方程为

$$P_{k+1} = (I - K_{k+1} H) P_{k+1/k} \tag{10}$$

A 为状态转移矩阵，H_1 为观测矩阵，w_k 为状态噪声 v_k 为观测噪声，均为高斯白噪声。

$P_{k+1/k}$ 和 X_k 的初始值分别为 $P_{k=1} = W$ 和 $X_1^t = \hat{X}_1^t$ ， $\hat{X}_1^t$ 为目标 t 出现的第一帧检测结果的状态向量，作为第一帧的估计值传递给第二帧进行滤波，其中五个变化值初始化为 0。从目标 t 出现第二帧开始，取当前帧的预测值 $\hat{X}_1^{t}{}'$ 和估计值 $\hat{X}_k^t$ 作为该帧图像的两个候选区域，连同 SSD 提取的候选区域一并提取池化特征。该帧检测结束后，将结果作为该帧滤波值送入下一帧滤波。当出现多个目标时则分别进行滤波，目标个数增加时增加相应个数的滤波器。此外，本文设定当目标连续十帧滤波值对应的候选区域没有作为检测结果时，取消该滤波器[10]。

改进后整体的检测算法框架流程如图 4 所示

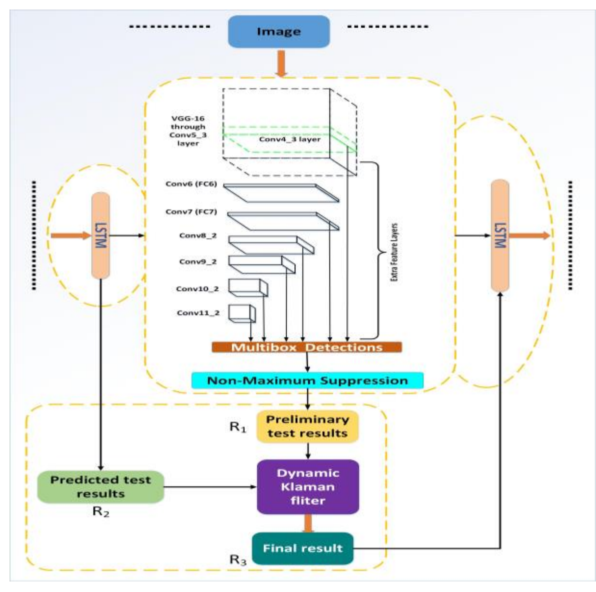


图 4　改进后算法整体框架

Fig. 4　Improved overall framework of the algorithm

算法流程如下：

a)将单帧视频图像输入 SSD Detector 结合 LSTM 网络传递的预测各层 feature map 进行目标检测识别，获得初检测结果 R_1；

b)通过 LSTM 网络传递获得当前帧的预测检测结果 R_2，通过 dynamic Klaman fliter，将初检测结果 R_1 和预测检测结果 R_2 结合起来，获得最终的检测识别结果 R_3；

c)将当前帧检测过程中产生的各层 feature map 以及检测结果 R_3 输入 LSTM 网络，对下一帧的检测结果进行预测指导。

3　实验结果与分析

3.1　实验的基础条件与数据集库

本文实验使用 Dell Precision R7910(AWR7910) 图形工作站，处理器为 Intel Xeon E5-2603 v2(1.8 GHz/10M)，采用 NVIDIA Quadro K620 GPU 加速运算。SSD 是基于深度学习框架 Caffe 来运行的。Caffe 支持 CPU 和 GPU 的并行运算，使得计算量庞大的深度学习得以在短期内完成。本文在 YFCC100M 收集的交通场景数据集(Web dataset)和 KITTI 数据集上进行了实验。KITTI 数据集由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创办，是目前国际上最大的自动驾驶场景下的计算机视觉算法评测数据集。选用 KITTI 数据集中第 1 个图片集 download left color images of object data set 和标注文件 download training labels of object data set ，其中 7481 张训练图片有标注信息，而测试图片没有。SSD 中训练脚本是基于 VOC 数据集格式的,、需要把 KITTI 数据集做成 PASCAL VOC 的格式。PASCAL VOC 数据集总共 20 个类别，本文为数据集设置 3 个类别 ‘Car’‘Cyclist’‘Pedestrian’。 YFCC100M 数据集包含将近 1 亿张图片以及摘要、标题和标签。为了更好地展示本文的方法，通过搜索关键词“行人”、“道路”和“车辆”从 YFCC100M 数据集收集了 1000 幅分辨率较高的测试图像。对于该数据集，使用至少 16 像素宽度和小于 50 %遮挡对所有目标进行注释。图像在较长的一侧被重新缩放到 2000 像素，以适合本文的 GPU 内存。实验中将所有的图像尺寸归一化为 320 x 320。

3.2　实验的参数设置

本文选择 SSD 系列中的 SSD512 进行改进，SSD512 提供了大、中、小三个规模的深度卷积神经网络模型，本文选取中等规模的 VGG_CNN_M_1024 模型作为基础模型，改动与目标类别数目相关的参数（原模型需要识别 20 类目标而本文只有 3 类）。

为了优化调参过程以及快速选取自适应池化纠正误差项的最佳值，制作了小样本数据集（200 张图像），在很大程度上节约了时间成本，提高了调参选值效率。在小样本的抽取上既要考虑总体的类别数，又要考虑每种类别占总体的比例大小，而概率抽样方法中的分层抽样能够很好地兼顾此两点[12]。因此按照该抽取规则，小样本数据集在一定的程度上可以代表原始数据集，通过小样本数据集训练所得的最优超参数在一定的程度上能够适应原始数据集。在不使用自适应阈值时，阈值设置为 0.7；将所有实验中经过非极大抑制留下的候选区域数量设置为 100（默认设置为 300）。其他设置保持默认不变，后续所有实验都在以上设置基础上进行。对于 LSTM，将 LSTM 展开到 10 个步骤，并按照 10 帧序列进行训练，通道宽度乘子 αδ=1，模型学习率为 0.003，其他参数与文献[7]一致。

3.3　评价指标

在多目标分类器的判别中， 设目标的种类数为 n 。对单种目标的判别仍然遵循每一种假设有两种结果的四种可能性，即

设 $D_i^j(j=1,2,\cdots,n)$ 表示一种目标 j 选择假设 H_i^j 为真， 任何二元假设实验问题中，作判别时要考虑四种可能性[13]：a）H_0^j 假设为真，判别为 D_0^j；b）H_0^j 假设为真，判别为 D_1^j；c）H_1^j 假设为真，判别为 D_0^j；d）H_1^j 假设为真，判别为 D_1^j。a）和 d）对目标 j 选择正确；b）称为第一类错误，叫作虚警（没有目标而识别为有目标）；c）称为第二类错误，叫做漏报（有目标而误判为没有目标）。除此之外，在多目标识别中将目标 D_i^j 识别为目标 $D_i^k(k=1,2,\cdots,n,k\neq j)$ 的错误判别。

设目标 z^j 在判别域 Z_0^j 和 Z_1^j 上的概率密度函数分别为 $f(z|H_0)$ 和 $f\left(z^j|H_1^j\right)$，则有

1）虚警率

$$P_f = \sum_{j=1}^{n} P\left(D_1^j|H_0^j\right) = \sum_{j=1}^{n} \int_{Z_1^j} f\left(z^j|H_0^j\right) dz \tag{11}$$

2）漏警率

$$P_m = \sum_{j=1}^{n} P\left(D_0^j \middle| H_1^j\right) = \sum_{j=1}^{n} \int_{Z_0^j} f\left(z^j \middle| H_1^j\right) dz \tag{12}$$

3）检测率

$$P_d = \sum_{j=1}^{n} P\left(D_1^j \middle| H_1^j\right) = \sum_{j=1}^{n} \int_{Z_1^j} f\left(z^j \middle| H_1^j\right) dz \tag{13}$$

4）误检率

$$P_e = \sum_{j=1}^{n} \sum_{k=1, j\neq k}^{n} P\left(D_1^j \middle| H_1^j\right) = \sum_{j=1}^{n} \sum_{k=1, j\neq k}^{n} \int_{Z_1^k} f\left(z^j \middle| H_1^j\right) dz \tag{14}$$

根据定义可知，虚警率、检测率、漏警率与误检率之和为 1。在实际计算时，首先计算识别率，再计算误报率、漏报率，对于剩余系统识别出来的而实际不存在的目标种类作计数来计算分类的虚警率。对于多目标识别中的虚警率应该计算一定时间段内积累的虚警率。对于数据集，采用求平均的方式来计算整体的虚警率、漏警率、检测率、误检率。

深度学习通过误差的反向传播来调整神经网络权值，达到建模的目的。反向传播迭代次数从几万次逐步增加到数十万次，直到训练误差趋于收敛为止。最后通过计算模型在测试集上的平均准确率（average precision，*AP*）和所有类别的平均准确率均值（mean AP，*m AP*）来评价模型的好坏。*AP* 从召回率和准确率两个角度衡量检测算法的准确性。*AP* 是评价深度检测模型准确性最直观的标准，可以用来分析单个类别的检测效果。*mAP* 是各个类别 *AP* 的平均值，*mAP* 越高表示模型在全部类别中检测的综合性能越高[10]。

3.4　实验设计

首先将各个策略与 SSD512 进行单独结合，进行相应的对比实验，表明各个策略的作用。然后将所有策略与 SSD512 结合，对最终的改进算法进行整体测评。用训练集训练原始 SSD512，将此模型记为 M0，在 M0 基础上加入 LSTM 递归神经网络，生成模型 M1；在 M1 基础上加入动态卡尔曼滤波策略，生成模型 M2，使用两数据库测试集对 M0、M1、M2 进行测试和对比。

另外选取了 Faster R-CNN、不需要预训练模型的 DSOD300[14]（deeply supervised object detector）检测框架和 YOLO 系列检测框架中的升级版 YOLOv2 544[15]，以及 SSD 的改进模型 DSSD[16]（deconvolutional single shot detector）作为深度学习对比算法，与 M2 对比 Web dataset 和 KITTI 数据集上的检测效果。对比检测框架算法使用作者发布的官方代码中的默认参数设置，与 M2 在相同训练集中进行训练。利用 Web dataset 和 KITTI 数据集中的测试集进行测试。

3.5　算法关键参数讨论

在 LSTM-SSD 体系结构中卷积层使用具有 384 通道的单个 LSTM。通过对 Bottleneck-LSTM 和 feature map 层应用附加卷积来获得最终边界框。

将所有四个 LSTM 门计算合并为单个卷积，因此 LSTM 计算 1 536 个通道的门但仅输出 384 个通道。为了解决过拟合问题，采用分两阶段的方法对网络进行训练。首先，在没有 LSTM 的情况下微调 SSD 网络；然后，保持基本网络中的权重，直到 Conv13 层(包括 Conv13 层)，并在剩余的训练中插入 LSTM 层。

在网络模型中的不同层之后放置单个 LSTM 层（α=1　）。表 1 证实了将 LSTM 放置在特征映射之后可获得识别性能的提高，其中放在 feature map1 层后提高效果最为明显，从而验证了本文关于在特征空间中添加时间感知对提高检测识别精度的有效性。

3.6　实验结果

实验结果如表 2、3 所示，分别对比了模型 M0、M1、M2 在 KITTI 和 WD 数据集上普通测试集的识别与检测效果。

表 1　LSTM 插入位置对识别率的影响

Table 1　Effect of LSTM insertion position on recognition rate

placed after	dataset	*AP*(%)			*mAP*(%)
		Person	Car	Cyclist	
baseline	KITTI	73.36	71.53	65.32	70.07
Conv3	KITTI	66.72	61.32	59.03	62.36
Conv13	KITTI	76.28	72.12	64.49	70.96
feature map1	KITTI	**77.21**	**75.08**	**68.62**	**73.64**
feature map2	KITTI	72.08	72.24	66.35	70.22
feature map3	KITTI	72.16	71.02	67.13	70.10
feature map4	KITTI	75.25	70.43	67.41	71.03
outputs	KITTI	74.86	72.19	66.92	71.32

表 2 各模型识别精度对比

Table 2　Comparison of recognition accuracy of each model

model	dataset	*AP*(%)			*mAP*(%)
		Person	Car	Cyclist	
M0	KITTI	73.36	71.53	65.32	70.07
	WD	71.59	69.63	62.75	67.99
M1	KITTI	85.18	79.35	74.69	79.14
	WD	72.52	70.45	64.83	69.27
M2	KITTI	88.42	81.73	74.38	81.51
	WD	74.92	72.34	65.63	70.96

表 3 各模型检测效果对比

Table 3　Comparison of test results of each model

model	dataset	P_f (%)	*Pm* (%)	*Pd* (%)	*Pe* (%)
M0	KITTI	20.21	19.34	41.32	19.13
	WD	19.25	21.38	38.83	20.54
M1	KITTI	16.31	16.29	50.84	16.56
	WD	18.17	19.49	43.45	18.89
M2	KITTI	9.53	11.69	64.25	14.53
	WD	16.24	15.19	51.16	17.41

对比表 2、3 中 M0 和 M2 检测结果，在 KITTI 数据集中，各类目标检测的 *AP* 提高了 9%~15%不等，*mAP* 提高了约 11.44%，虚警率降低 10.68%，检测率提高 22.93%，漏警率降低 7.65%，误检率降低 4.6%；在 WD 数据集中，各类目标检测的 AP 提高了 1~3%不等，mAP 提高了约 2.97%，虚警率降低 3.01%，检测率提高 12.33%，漏警率降低 6.19%，误检率降低 3.13%。M2 模型是在 M0 基础上加入时间感知 LSTM 网络和动态卡尔曼滤波策略训练得到的，通过在两个数据库上的测试结果与 M0 对比可以发现，M2 相较于 M0，多目标的检测率得到了较大提高，多目标检测的虚警率和漏警率降低明显，对各目标的识别精度和平均识别精度同样获得了较大的提高。而且，由于 WD 数据集是静态图像数据集，时空上下文策略无法生效，改进效果不如在视频数据集 KITTI 上的效果明显。表明基于时间感知特征映射的移动视频目标检测改进策略能够有效降低 SSD512 对视频中多目标检测的漏警率和虚警率，较大地提高目标识别精度。各项指标提升明显，表明本文策略总体对于弥补 SSD512 缺陷的有效性。

对比表 2、3 中 M1 和 M2 检测结果，在 KITTI 数据集中，各类目标检测的 *AP* 提高了 1~4%不等，*mAP* 提高了约 2.67%，虚警率降低 6.78%，检测率提高 13.41%，漏警率降低 4.6%，误检率降低 2.03%；在 WD 数据集中，各类目标检测的 *AP* 提高了 1~3%不等，*mAP* 提高了约 1.69%，虚警

率降低 1.93%，检测率提高 7.71%，漏警率降低 4.3%，误检率降低 1.48%。M2 模型是在 M1 基础上加入动态卡尔曼滤波跟踪策略训练得到的，通过在两个数据库上的测试结果与 M1 对比可以发现，M2 相较于 M1，对多目标的检测率得到了较大提高，多目标检测的虚警率和漏警率降低明显，表明动态卡尔曼滤波跟踪策略有效增强了了对遮挡等强干扰目标检测的鲁棒性。

为了进一步验证 M2 模型已经学习到视频的时间连续性，对于遮挡等干扰具有较强的鲁棒性，在 KITTI 视频数据集中单帧图像上创建人工遮挡来进行测试。对于图像中每个目标的真实检测框，按照目标遮挡率 $p_z \in (0,1]$，来设计人工遮挡。对于尺寸为 $H\times W$ 的目标真实检测框，在检测框内随机选择一块尺寸为 $p_z \cdot H \times p_z \cdot W$ 的区域，将该区域内的所有像素值都取为 0，这样就构成了人工遮挡。将 KITTI 视频数据集中普通测试集每隔 50 帧随机挑选目标构造人工遮挡，构造抗遮挡鲁棒性测试集，M0、M3 在这个测试集上进行测试，取目标遮挡率分别为 *Pz*=0.25、*Pz*=0.5、*Pz*=0.75、*Pz*=0.1，测试结果如表 4 所示

表 4　M2 抗遮挡干扰效果验证

Table 4　M2 anti-occlusion interference verification

Model	Evaluation metric	Pz=0.25	Pz=0.5	Pz=0.75	Pz=0.1
M0	mAP（%）	53.36	41.24	22.15	12.89
	Pd（%）	33.58	21.56	12.33	4.25
M2	mAP（%）	74.28	66.82	59.79	51.58
	Pd（%）	60.35	55.62	51.16	42.39

由表 4 对比 M0、M2 在不同目标遮挡率下的 *mAP*、检测率 P_d，可以发现本文的方法在这种遮挡噪声数据上优于单帧 SSD 方法，表明网络已经学习到视频的时间连续性，并且使用时间线索来实现对遮挡噪声的鲁棒性。

利用 Web Dataset 和 KITTI 数据集中的普通测试集进行测试。检测识别效果如表 5 所示，其中 *FPS* 代表算法运行的速度、帧率

表 5　各检测算法检测识别效果对比

Table 5　Comparison of detection and recognition effects of each detection algorithm

method	dataset	*AP*(%)			*mAP*(%)	*Pd* (%)	*FPS*
		Person	Car	Cyclist			
Faster R-CNN	KITTI	83.26	74.13	75.42	77.61	45.22	13.15
	WD	81.49	71.33	68.65	73.82	36.63	11.64
DSOD300	KITTI	77.43	72.26	68.38	72.69	58.68	58.23
	WD	70.73	69.39	67.04	69.05	52.32	50.35
DSSD513	KITTI	75.46	69.53	68.34	71.11	59.42	46.34
	WD	72.19	68.83	66.45	69.16	49.79	39.38
YOLOv2 544	KITTI	79.43	71.25	67.32	72.66	60.82	56.74
	WD	73.29	69.63	68.85	70.59	54.86	49.28
M2	KITTI	88.42	81.73	74.38	81.51	64.25	42.56
	WD	74.92	72.34	65.63	70.96	51.16	32.83

对比表 5 中 M2 和 其他深度学习对比算法检测结果，在 KITTI 数据集中，各类目标识别的 *AP* 提高了 5~16%不等，*mAP* 提高了约 4~10%不等，检测率提高 4~19%；在 WD 数据集中，相比于 DSSD513、Faster R-CNN 检测率分别提高 1.37%、15.53%。虽然检测识别速率比不上 DSOD300、DSSD513、YOLOv2 544 等检测算法，但是 *FPS* 也能达到 43 帧/s，基本能够满足实时性的要求。

4　结束语

针对现有基于大数据和深度学习的目标检测框架难以实现在低功耗移动和嵌入式设备上实时进行视频目标检测的问题，改进了基于深度学习的目标检测框架 SSD，提出一种改进的多目标检测框架 LSTM-SSD，将其专用于交通场景视频多目标检测。实验表明，改进后的在应对弱小目标、多目标、杂乱背景、光照变化、模糊、大面积遮挡等检测难度较大的情况时，均能获得较好的效果，实现了算法精度与运行速率的平衡，为深度学习在特定目标检测的应用提供了实例和新的思路。但是算法的处理效率距离工程实际应用的需求仍然有差距，且对低分辨率小目标的识别效果并不理想，后期如何降低运算量提高算法的实时性和针对低分辨率弱小目标的检测和识别将是主要的研究方向。

参考文献：

[1] 迟晓君，孟庆春，陈鹏．基于最小风险的 Bayes 决策方法在交通检测中的应用 [J]．计算机应用研究, 2005, 22(12): 204-205. (Chi Xiaojun, Meng Qingchun, Chen Peng. Application of Bayesian decision-making method based on minimum risk in traffic detection [J]. Application Research of Computers, 2005, 22(12): 204-205.)

[2] 于凯，贾磊，陈宇强．深度学习的昨天，今天和明天 [J]．计算机研究与发展，2013, 50(9): 1799-1804. (Yu K, Jia L, Chen Y Q. Deep learning: yesterday, today, and tomorrow. Journal of Computer Research and Development, 2013, 50(9): 1799-1804.)

[3] Liu Wei, *et al.* SSD: single shot multibox detector [C]//Proc of European Conference on Computer Vision. Cham:Springer,2016: 21-37.

[4] Ren Shaoqing, *et al.* Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[5] Redmon, Joseph, *et al.* "You only look once: unified, real-time object detection[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washingtom,DC: IEEE Computer Society, 2016: 779-788.

[6] Lin Tsuang Yi, *et al.* Feature Pyramid Networks for Object Detection [J]. arXiv preprint arXiv: 1612. 03144, 2016.

[7] Liu Mason, *et al.* Mobile Video Object Detection with Temporally-Aware Feature Maps [J]. arXiv preprint arXiv: 1711. 06368, 2017.

[8] 赵鹏，刘杨，刘慧婷，等．基于深度卷积-递归神经网络的手绘草图识别方法 [J]．计算机辅助设计与图形学学报, 2018(2) . (Zhao Peng, Liu Yang, Liu Huiting, *et al.* A Hand-drawn sketch recognition method based on depth convolution-recursive neural network [J]. Journal of Computer-Aided Design & Computer Graphics, 2018(2) .)

[9] 王多民，刘淑芬．一种使用 log 函数的新型修正激活单元 LogReLU [J]．吉林大学学报:理学版 ，2017, 55(3): 617-622. (Wang Duomin, Liu Shufen. A new modified activation unit LogReLU using log function [J]. Journal of Jilin University (Science Edition) , 2017, 55(3): 617-622.)

[10] 冯小雨，梅 卫，胡大帅．基于改进 Faster R-CNN 的空中目标检测

[J]. 光学学报, 2018, 38(6): 0615004. (Feng Xiaoyu, Mei Wei, Hu Dashuai. Aerial target detection based on improved Faster R-CNN [J]. Acta Optica Sinica, 2018, 38(6): 0615004.)

[11] 李成龙, 钟凡, 马昕, 等. 基于卡尔曼滤波和随机回归森林的实时头部姿态估计 [J]. 计算机辅助设计与图形学学报, 2017, 29(12): 2309-2316. (Li Chenglong, Zhong Fan, Ma Wei, *et al*. Real-time head pose estimation based on Kalman filtering and stochastic regression forest [J]. Journal of Computer-Aided Design & Computer Graphics, 2017, 29(12): 2309-2316.)

[12] 胡聪, 屈瑾瑾, 许川佩, 等. 基于自适应池化的神经网络的服装图像识别[J]. 计算机应用, 2018, 38(8): 2211-2217.. (Hu Cong, Qu Wei, Xu Chuanpei, Zhu Aijun. Apparel image recognition based on adaptive pooling neural network [J]. Computer Application , 2018, 38(8): 2211-2217.)

[13] 马春庭, 郑坚, 陈东根, 等. 地面战场侦察系统多目标识别的评价指标 [J]. 探测与控制学报, 2006, 28(1): 6-9. (Ma Chunting, Zheng Jian, Chen Donggen, *et al.* Evaluation index of multi-target recognition for ground battlefield reconnaissance system [J]. Journal of Detection & Control, 2006, 28(1): 6-9.)

[14] Shen Zhiqiang, *et al*. DSOD: learning deeply supervised object detectors from scratch [C]//Proc of IEEE International Conference on Computer Vision. IEEE Computer Society, 2017: 1937-1945.

[15] Zhang Jianming, *et al.* A real-time chinese traffic sign detection algorithm based on modified YOLOv2 [J]. Algorithms, 2017, 10(4): 127.

[16] Fu Cheng Yang, *et al*. DSSD: deconvolutional single shot detector [J]. arXiv preprint arXiv: 1705. 09587, 2017.

[17] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述 [J]. 计算机学报, 2017, 40(6): 1229-1251. (Zhou Feiyan, Jin Linpeng, Dong Jun. A review of convolutional neural networks [J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251.)

[18] 常亮, 邓小明, 周明全, 等. 图像理解中的卷积神经网络 [J]. 自动化学报, 2016, 42(9): 1300-1312. (Chang Liang, Deng Xiao-ming, Zhou Ming-quan, *et al.* Convolutional neural networks in image comprehension [J]. Acta Automatica Sinica, 2016, 42(9): 1300-1312.)